REVIEW

New sequencing technologies, the development of genomics tools, and their applications in evolutionary arachnology

Michael S. Brewer[1,2], **Darko D. Cotoras**[3], **Peter J. P. Croucher**[1,2] and **Rosemary G. Gillespie**[1,2]: [1]Department of Environmental Science, Policy, and Management, University of California, Berkeley, California 94720 USA. E-mail: michaelsbrewer@gmail.com; [2]Essig Museum of Entomology, University of California, Berkeley, California 94720 USA; [3]Department of Integrative Biology, University of California, Berkeley, California 94720 USA

Abstract. Molecular genetic tools have been a boon to arachnologists for decades and used to study many unique aspects of arachnid biology including genomics, phylogenetics, population genetics, and biogeography. These tools have evolved over time and now provide myriad methods for exploring evolutionary questions. Early tools, while still useful under the proper circumstances, are giving way to a new generation of DNA sequencing technologies. These new platforms yield impressive amounts of data at a fraction of the cost of traditional techniques. Herein, we discuss the history and future of molecular evolutionary arachnology in terms of available genetic/genomic tools and their potential applications, strengths, weaknesses, and relative costs. Next-generation sequencing (NGS) platforms are varied in their methods and potential uses, making high-throughput sequencing studies focusing on a wide array of questions tractable. To date, relatively few studies have employed NGS technologies using arachnids, but many could benefit from using them. Because no model species exist within the class Arachnida, we have a limited understanding of arachnid genomics. With the ever-advancing nature of sequencing technologies and bioinformatics, arachnologists can relatively easily implement NGS studies to bridge the gaps in our understanding and open avenues for deeper and more powerful experiments. To this end, we discuss examples of applications of NGS technologies focusing on arachnid taxa. Despite the allure of acquiring massive quantities of sequence data, we should recognize the limitations of existing NGS technologies and not forsake pre-NGS methods when these technologies could adequately address our questions.

Keywords: Next-generation sequencing, genome, transcriptome, phylogeny, population genetics, genomics, adaptation, selection

TABLE OF CONTENTS

1. INTRODUCTION

Since Linnaeus and before, scientists have sought to put order into the diversity of life, the thirst for information increasing with the recognition of the role of evolutionary processes in shaping that diversity. The advent of molecular techniques in the 1980s introduced a huge diversity of novel markers for assessment of phylogenetic affinities. Moreover, with the growth of the human genome project, the potential use of vast numbers of genes across the genome was soon recognized (Jones 1991). Analytical tools were developed that could use the almost limitless data to address questions ranging from historical and recent demographic and migratory patterns to identifying signatures of recent natural selection (Nielsen 2010; Rasmussen et al. 2011; Lohmueller 2011). Here, we examine where the field of arachnology stands within the genomic revolution.

In recent years, advances in sequencing technology have led to great increases in genomic resources for many non-model species. The arthropod class Arachnida, encompassing over 100,000 nominal species classified into 12–13 traditional orders (Krantz and Walter 2009; Blick and Harvey 2011), comprises a diverse array of taxa that serve key functions in terrestrial ecosystems as important predators and decomposers. Mueh of their diversity is unique to particular arachnid taxa, including complex silk production (spiders and mites), venom composition (spiders, scorpions, and pseudoscorpions) and detoxification of plant compounds (Grbić et al. 2011), and has long been of particular interest to researchers. As an example, spiders have been used to study behavior (reviewed by Herberstein and Hebets 2013), development (e.g., Kanayama et al. 2010; Wolff and Hilbrant 2011; Mittmann and Wolff 2012), sexual selection (e.g., Kuntner et al. 2009; Su et al. 2011), genetics (reviewed by Goodacre 2013), evolutionary ecology (reviewed by Moya-Laraño et al. 2013) and biogeography (reviewed by Gillespie 2013), among other fields. However, relationships within and among arachnid taxa are, in many cases, not presently resolved (Giribet and Edgecombe 2013), and a paucity of genomic resources has hindered efforts in various fields of arachnology. Molecular teehniques have been successfully employed to investigate a number of these issues, but some problems require data at a scale heretofore unavailable to arachnologists.

Here we first briefly review how more "traditional" molecular tools have been applied to arachnid biology and then go on to discuss emerging next-generation sequencing ("NGS") applications and their potential impact within the field. Instead of deeply discussing each area of arachnology and thoroughly reviewing the literature, we provide a brief background with select citations and proceed to describe some ways in which the rapidly multiplying set of NGS tools may be used.

2. "TRADITIONAL" MARKERS OF NUCLEAR GENETIC VARIATION

Prior to genomic tools, multiple techniques were used for examining variation in nuclear DNA and hence to assess geographic and population structure, the most widely used being randomly amplified polymorphic DNA (RAPDs), restriction fragment-length polymorphisms (RFLPs), and satellite and microsatellite DNA.

2.1. Allozyme electrophoresis.—Allozyme electrophoresis has proven very useful for the analysis of geographic structure of arachnids. Some early studies focused specifically on population structure (Porter & Jakob 1990; Steiner et al. 1992; Smith & Engel 1994; Hudson & Adams 1996; Smith & Hagen 1996; Boulton et al. 1998), but allozymes have also been used to examine questions of relatedness among colonies of social spiders (Johannesen et al. 1998; Johannesen & Lubin 1999, 2001; Johannesen & Veith 2001; Evans & Goodisman 2002; Yip et al. 2012), paternity (Schafer & Uhl 2002), species boundaries and speciation (Piel & Nutt 2000; Ramirez & Chi 2004), dispersal (Pedersen & Loeschcke 2001; Schafer et al. 2001), the effects of forest fragmentation, whether natural (Vandergast et al. 2004) or manmade (Ramirez & Haakonsen 1999; Gurdebeke et al. 2000), and to estimate selection on color polymorphisms (Tso et al. 2002; Oxford 2005; Oxford & Gunnarsson 2006; Croucher et al. 2012) as well as patterns of diversification within rapidly diversifying lineages (Pons & Gillespie 2004; Baert et al. 2008; De Busschere et al. 2010).

The primary limitations of allozyme electrophoresis are: a) Organisms must generally be alive or deep-frozen before use; b) when bands co-migrate, they are assumed to be homologous; c) only a very small subset of the genetic variation at a given locus is revealed; and d) it is not possible to distinguish ancestry and descent among different alleles. The technique is inexpensive, fast, and can give insight into multiple loci and so is useful for addressing questions of geographic structure.

2.2. Satellite & microsatellite DNA.—Tandem repeats include three subclasses: satellites, minisatellites and microsatellites. Satellites range in size from 100 kb to over 1 Mb with repeat units of ca. 100–200 bp; most are located at the eentromere. Minisatellites range from 1 kb to 20 kb in size with shorter repeats (9–80 bp), while microsatellites (also known as short tandem repeats, STR), are repeats of sequences less than about five base pairs in length (an arbitrary cutoff). Among spiders, satellite DNA has proven very useful in assessing relationships among species within a radiation of spiders in Hawaii; this was because the tandem-arranged units show a high intraspecific sequence identity due to concerted evolution (Pons & Gillespie 2003, 2004). As a result, the length of the branches and corresponding support were much greater for satellite DNA than for mtDNA sequence data.

Microsatellites are repeated short sequences of DNA that occur throughout the genomes of many organisms, including spiders. Because repeat units are readily added to or lost from microsatellite DNA, the sequence length of these regions evolves rapidly. Microsatellites offer a valuable pool of genetic variation that has proven very useful for looking at paternity and relatedness among spiders, including social species (Ji et al. 2004; Bilde et al. 2009; Duncan et al. 2010), as well as understanding geographic structure between closely related populations (Rutten et al. 2001; Reed et al. 2007, 2011; Krehenwinkel & Tautz 2013; Parmakelis et al. 2013). However, compared to many other fields, the development and application of microsatellites in spider ecology and evolution has been limited. A potential cause for the paucity of microsatellite studies in arachnids is the apparent difficulty in finding reliable loci. However, the low % GC (percentage of guanine and cytosine residues in DNA sequences) in some lineages, as discussed below, could potentially play a role.

2.3. Random amplified polymorphic DNA (RAPD).—In the RAPD procedure, a single nucleotide primer (8–10 base pairs long) is used to amplify random sections of nuclear DNA, with differences in band sizes being used to provide information on relationships. The method has been used in spiders (e.g., A'Hara et al. 1998; Gurdebeke et al. 2003). However, although the approach provides a lot of variability, RAPDs suffer from poor repeatability, lack of codominance, and the possibility of non-heritable or non-homologous bands.

2.4. Restriction fragment length polymorphisms (RFLP).—For generating RFLPs, regions of nuclear DNA isolated through PCR or other means can be digested with restriction enzymes that cut samples of homologous DNA at specific four- or six-base sequences, differences arising from the locations of restriction enzyme sites. This technique could, compared to other technologies of the time, exploit an enormous amount of genetic variation. However, although used in mites (e.g., Osakabi & Sakagami 1994), it was never an important technique in other arachnid groups.

2.5. Amplified fragment length polymorphisms (AFLP).—AFLPs use restriction enzymes to digest genomic DNA, with the fragments then amplified and separated, providing markers across many loci that are highly variable and are also reproducible. Like RAPDs, however, they are also both anonymous and dominant and may produce non-homologous bands. They have been used in studies of geographic structure among spider populations (Jung et al. 2006; Lambeets et al. 2010; Croucher et al. 2011a, b), where they provided fine resolution of population differentiation and subdivision. They have also been used in assessments of inbreeding and sociality (Bilde et al. 2005).

3. SEQUENCING METHODS

3.1. Sanger sequencing of mitochondrial DNA.—Mitochondrial DNA, notably the cytochrome oxidase 1 (CO1), NADH dehydrogenase 1 (ND1) and 16S rRNA genes (Agnarsson et al. 2013) proved particularly useful in the earliest studies of biogeography and species differentiation in spiders (Gillespie et al. 1994; Hedin 1997a, b; Johannesen et al. 2002). The reason for this is simply because of the abundance of mitochondrial DNA relative to nuclear DNA, making it much easier to amplify. However, problems with mtDNA that affect recently evolving lineages include the lack of recombination, as a result of which it behaves as a single locus, making it of limited value for analytical approaches requiring multiple loci. This makes its use in species delimitation particularly problematic (Hamilton et al. 2014). Moreover, the haploid nature of mitochondrial DNA means that the marker is more sensitive to small population sizes than is nuclear DNA, and the maternal inheritance means that biases in movement between sexes cannot be recovered. For this reason, recent studies that have used mtDNA sequences have generally included various nuclear markers (e.g., Vandergast et al. 2004; Starrett & Hedin 2007; Croucher et al. 2011a & b, 2012; Satler et al. 2013). Mitochondrial DNA has been applied to questions at deeper phylogenetic levels where the microevolutionary problems mentioned above are less severe. However, the rapid evolution of the marker means that it tends to become saturated rather quickly (Brewer et al. 2013).

3.2. Sanger sequencing of nuclear DNA.—Because of the issues of amplification, the most reliable, and hence useful, nuclear genes have tended to be those that occur in multiple copies such as Histone 3, and the ribosomal 18S and 28S genes, and these have been of particular impact in the realm of phylogenetic reconstruction (reviewed by Agnarsson et al. 2013 and Giribet and Edgecombe 2013). At the population-species level, attention has focused on nuclear introns—noncoding sequences within nuclear genes, as these are not subject to the same selective constraints as exons and tend to evolve faster (Garb & Gillespie 2009; Hedin et al 2010). ITS (internal transcribed spacer) regions within the ribosomal RNA genes can frequently provide sufficient variability at shallow levels (Hormiga et al. 2003), though paralogy can often make the identification of homologous DNA difficult or impossible. Indeed, the problem of generating multilocus (nuclear) data has remained. Thus, researchers have looked increasingly toward modern, high-throughput (or "next-generation") sequencing technologies as a potential means to generate large amounts of multilocus data by increasing the amount of data per monetary cost by orders of magnitude.

3.3. Sanger sequencing versus NGS methods.—In the past, DNA sequence data were primarily collected using dideoxyribonucleotide (ddNTP) termination methods (i.e., Sanger sequencing). This approach provides long, high-quality sequences but suffers from a number of limitations (Table 1), including the ability to sequence only a single locus per reaction. In addition, reactions typically require taxon (or even population) specific oligonucleotide "primers"—short fragments of DNA (ca. 20–25 bp) of known sequence for polymerase chain reaction amplification and the sequencing reaction. Lastly, the cost of collecting data is much higher than in NGS approaches. Sanger sequencing methods are quite scalable in that one can easily obtain data for a single locus to hundreds of loci with a concomitant change in cost, but one cannot sequence massive amounts of genomic data from numerous specimens in a cost effective (in terms of time and money) manner.

In contrast to Sanger sequencing, NGS techniques provide vastly larger quantities of data much faster and for far less money. NGS approaches achieve this in two ways. The first involves ligating or attaching adaptors (synthesized DNA strands of known sequence) to the ends of fragments of target DNA. These adaptors allow identical sets of PCR or sequencing primers to be used for all the DNA fragments (a form of "shotgun" sequencing). From an operational point of view, the major differences between the various NGS approaches are in the size of the DNA fragment (the "insert") and the number of base pairs of sequence data that can be recovered from the end of the fragment. The second way that NGS approaches dramatically reduce costs is by miniaturization and parallelization—millions of sequencing reactions take place in small reaction chambers or flow cells (Shendure & Ji 2008). Consequently, high-throughput methods can sequence numerous DNA fragments concurrently and in a single reaction/run with full-length sequences typically being assembled after the fact.

Table 1.—Summary of the strengths, weaknesses, starting material and applications of select molecular data sources. Pre- and post-next-generation sequencing molecular data sources discussed here are listed. Several positives and negatives are given for each technique, along with differences in starting material required. Additionally, several historical and potential applications in arachnology are provided for each method.

Method	Raw sequence output	Cost per Mb	Positives	Negatives	Potential in arachnids
Traditional Sanger sequencing	1.9–84 Kb	$2,400	Long, high quality reads; scalability	Very high cost per Mb	Traditional phylogenetics, population genetics and studies of few genes
454 pyrosequencing	0.7 Gb	$10	Long reads	Problems with homopolymers; high cost per Mb (for next-gen technology)	Genomes, transcriptome and microbiome studies
Illumina (Solexa) sequencing	600 Gb (HiSeq 2000)	$0.05–$0.15	High output; relatively low cost; widely used and supported	Short reads	Genome, transcriptome, massively barcoded amplicon and microbiomes studies
SOLiD sequencing	120 Gb	$0.13	Highly accurate	Short reads; not as widely supported	Genome, transcriptome, epigenetic and resequencing studies
Ion Torrent sequencing	20 Mb–1 Gb	$1	Scalability	Short reads	Transcriptome and barcoded amplicon studies
MiSeq sequencing	1.5–2 Gb	$0.50	Longer reads and smaller scale than older Illumina technologies	Output too low for some applications	Transcriptome, barcoded amplicon and microbiome studies
Single molecule real time (SMRT)	400 Mb	$0.75–$1.50	Very long reads	High error rate; best combined with other technologies	Genome and transcriptome studies

3.4. NGS versatility in sequencing targets.—The "shotgun" nature of NGS, using ligated universal adaptors, gives these approaches tremendous versatility in terms of what can be sequenced. This versatility facilitates genomic data collection for organisms about which little or no prior genetic information is known. Sequenced DNA targets can therefore theoretically consist of any source of DNA from total genomic DNA for genome sequencing (see Section 5) to cDNA (derived from total RNA by reverse transcription of expressed mRNA) (Mortazavi et al. 2008) from whole organisms or specific tissues that may have experienced different "treatments" ("transcriptomics" and "differential expression"). RNAseq libraries used for transcriptome sequencing target only the transcribed portion of the genome and are therefore a type of "reduced representation library" (RRL) (Van Tassell et al. 2008), and we discuss this approach along with other RRL approaches such as Exon Capture (Bi et al. 2012) and RADseq (Miller et al. 2007) below. RRL approaches target specific loci and produces fewer unique reads per individual than whole genome libraries and on certain NGS platforms, especially Illumina (see below), may produce highly redundant amounts of data per individual. This has resulted in methodologies that multiplex numerous individuals using small, unique oligonucleotide indices (i.e., tags or barcodes). These are typically incorporated into the adaptors and allow the sequences from each individual to be post-hoc sorted computationally. Barcoding therefore permits NGS approaches to act as high-throughput variant detection and genotyping platforms (e.g., Dahl et al. 2007; Meyer et al. 2008). Barcoding also comes into its own when the DNA targets originate from amplicons generated by traditional PCR approaches. Amplicons might be derived from long-range PCR of mitochondrial genomes, for example, or from standard molecular markers such as bacterial 16S, fungal 18S, or metazoan COI. Such massive barcoding approaches with amplicon sequencing are permitting community-wide metagenomic/microbiome analyses (Amaral-Zettler et al. 2009; Gloor et al. 2010; Caporaso et al. 2012) and large-scale phylogenetic studies (see below).

4. NEXT-GENERATION TECHNOLOGY PLATFORMS

Modern, high-throughput (or "next-generation") sequencing technologies have made many questions more tractable by increasing the amount of data per monetary cost by orders of magnitude. Although a number of NGS sequencing techniques have a steep learning curve (in terms of both wet-laboratory work and bioinformatics), much can be outsourced, and myriad computational resources (many of which may be used free of charge) are readily available. Some of these techniques are widely used, while others are still more limited in their availability. All such methods have inherent strengths and weaknesses that can be leveraged to address a wide range of questions. As more molecular data are collected for arachnid taxa, these groups may begin to approach "model" organism status in terms of a foundational understanding of genetics. This will allow more in-depth studies, using complex and powerful genetic and genomic techniques, to understand the basis of arachnid-specific traits.

4.1. Second-generation NGS technologies.—The basic logic behind several NGS technologies has been reviewed in recent works (Liu et al. 2012; McCormack et al. 2013; Quail et al. 2012). Therefore, we do not delve into specifics of the approaches, instead choosing to highlight strengths and weaknesses of the platforms. Many of the points below are summarized in Tables 1 and 2.

The first mainstream high-throughput technology was the Roche 454 system. This method relies on pyrosequencing

Table 2.—Comparison of DNA sequencing technologies with an emphasis on uses in arachnology. Several aspects of common sequencing technology are compared. The raw sequence output is highly variable between platform and shows the potential scalability provided by using different sequencing platforms. The cost per one million base pairs of data between methods also differs substantially and must be considered when attempting a study requiring sequencing. To help with choosing between platforms, we provide select positives and negatives for each. Finally, some potential high-level applications in arachnology are given.

Data source	Positives	Negatives	Starting material	Applications in arachnids
Pre next-gen sequencing				
Allozyme electrophoresis	Ease of use; widely used in arachnology	Requires fresh or frozen material; uncertain homology	Fresh or frozen specimens	Population genetics and biogeography
Variable nucleotide tandem repeats (VNTR; satellites)	More certain homology; widely used outside of spiders (many analytical tools)	Difficult to design; may not work between even closely related taxa	gDNA	Paternity, population genetics and biogeography
Random amplified polymorphic DNA (RAPD)	Easier to implement than satellite techniques	Lack of repeatability; codominance; uncertain homology	gDNA	Population genetics and biogeography
Restriction fragment length polymorphisms (RFLP)	Ease of use	Fragments of same size may not be homologous; not widely used in arachnology	gDNA	Population genetics and biogeography
Amplified fragment length polymorphisms (AFLP)	Widely used	Anonymous and dominant; uneertain homology	gDNA	Assessments of inbreeding; population genetics and biogeography
Termination (Sanger) sequencing	Sequence data; homology easier to infer; highly scalable compared to other non-NGS techniques	Much more expensive than NGS at a per base level; taxon/ population specific primers needed	gDNA	Assessments of inbreeding, population genetics, biogeography, phylogenetics and single gene studies
Post next-gen sequencing				
Genome sequencing	Full sequence data, creates foundation for many future studies	Costly, difficult, and unnecessary for many projects	High quality, high quantity gDNA	Genome studies, genetic mapping and developing model organisms
Transcriptome sequencing	Easy to sequence and serves a wide range of projects	Costly and biased towards coding regions of genome	High quality RNA from fresh or frozen specimens	Studying coding sequencing, differential expression, identifying isoforms, evolution of gene families, functional genes, and deep phylogenomics
RAD Tags	Powerful for genetic mapping, population genetics and phylogeography	Requires large amount of high-quality DNA	High quality, high quantity gDNA	Phylogeography, population genetics, species-level phylogeny and genomic mapping
Target Capture	Targets portions of the genome for wide array of projects	Requires additional genomic information	gDNA from fresh or preserved specimens (quality depends on application)	Studies of specific regions of the genome, functional genes, population genetics, species-level phylogeny and phylogeography
Anchored enrichment	Orthology certainty and phylogenetics at various taxonomic levels	Not developed in all groups	gDNA from fresh or preserved specimens (quality depends on application)	Phylogenetics of various taxonomic depths

chemistry to obtain millions of unique reads. Roche's 454 approach provides relatively long reads (~700 bp) at a much lower cost (~$10/Mb) than traditional sequencing methods (~$2,400/Mb). The 454 sequencing technology, although expensive per base of data in comparison to other NGS methods and yielding fewer unique sequence reads, is still widely used in genome and transcriptome sequencing and metagenomics because of the relatively long length of individual reads. However, other techniques (namely Illumina technologies, see below) can now serve many of the same functions as 454 sequencing at a much lower cost while providing many more unique sequence reads.

The second NGS platform to become widely used was Applied Biosystems sequencing by oligo ligation detection (SOLiD). This method uses the ligation of short probes to the template DNA. Each probe's extension relies on two-base matches, yielding highly accurate results at a lower cost-point (~$0.13/Mb) and in higher quantities than 454 sequencing (see below). A minor downside to the SOLiD sequencing platform is that the output is in a format unlike other technologies and requires computationally expensive algorithms to assemble. Nonetheless, this method can be used to efficiently study genomes, transcriptomes, and epigenetics (i.e., non-genetic modifications of the DNA sequence that affect expression such as methylation of CpG "islands", areas of the genome containing high frequencies of cytosine and guanine residues).

The last of the commonly used second-generation technologies, and the most frequently used NGS platform, is the Illumina system. Illumina chemistry relies on fixed flow-cell binding site oligonucleotides and complementary adaptors that also contain sequencing primer sites, and that are ligated to the DNA fragments to be sequenced. This technology yields a vast quantity of raw data (600 Gb) for relatively low cost ($0.05 – $0.15/Mb), and much effort has gone into developing novel ways of applying the method to a wide array of studies. These range from tweaking protocols to creating new algorithms and software for analyses. The main downside of the Illumina technology is that the sequence reads are relatively short. Early versions of the platform yielded reads that were only 36 bp in length, but read length as well as throughput continue to increase for all the NGS technologies, and the Illumina platform, for example, although still short in comparison to Sanger and 454 approaches, can now generate reads in excess of 150 bp. Short reads lead to complications in genome sequence assembly efforts and in community/microbiome sampling where assembly of sequences from a mixed pool of taxa is problematic. Fortunately, several approaches have been developed to address these problems including combining Illumina data with other sources, using large insert sizes for scaffolding, and overlapping reads for metagenomic amplicon sequencing (Masella et al. 2012). Therefore, Illumina sequencing is often used in genome sequencing efforts, transcriptomics, community sampling, resequencing, target enrichment, and many more techniques.

4.2. Compact personal genome sequencers.—In an attempt to down-scale high-throughput sequencing technologies to provide a more manageable amount of data for less money, "personal genome sequencers" have been developed. These machines are less expensive to buy, use, and maintain; hence individual labs may realistically own these machines for smaller scale and exploratory sequencing experiments. The first of these, the personal genome machine (PGM), was released by Ion Torrent (Rothberg et al. 2012). This platform is unique in its scalability. Sequencing takes place on individual disposable chips that can collect variable amounts of sequence data for differing levels of cost. This method is used for small genome sequencing (e.g., organellar or prokaryotic genomes) and transcriptomes. The Illumina MiSeq is similar in application to the larger Illumina platform, but at a smaller scale. Sequencing and data analysis are integrated into a single machine and can yield analyzed data in a single day. This method is commonly applied in highly multiplexed amplicon sequencing, small genome sequencing, microbial community analysis (Caporaso et al. 2012) and for the identification of transcription factors (i.e., ChiP-Seq).

4.3. Third-generation NGS technologies.—The newest high-throughput sequencing platforms, or single molecule sequencing, include two main technologies—single molecule real-time (SMRT) sequencing by Pacific Biosciences (PacBio) and the unreleased Nanopore platform (Oxford Nanopore Technologies, Oxford, UK). These methods are characterized by two main features: 1) no PCR prior to sequencing (limiting artifacts) and 2) sequences are recorded in real-time (i.e., during the polymerase reaction or depolymerization). These methods can each yield very long reads (>5 Kb and up to 13–14 Kb for PacBio) making them useful in de novo genome sequencing efforts. Short reads from the PacBio SMRT technology can be highly accurate since the platform has the ability to resequence the circularized molecule repeatedly until base confidences are high; however, long reads have a very high endemic error rate (ca. 15%). Approaches are being developed to correct the SMRT data using large quantities of accurate but short-read Illumina data (English et al. 2012; Koren et al. 2012). The Nanopore technology is currently not widely available, so much is still unknown concerning its performance. Moreover, although SMRT methods provide much longer reads than earlier NGS approaches with considerable simplification of library preparation, neither is currently well supported by common NGS bioinformatics tools.

5. ARACHNID GENOME EFFORTS

NGS technologies allow the sequencing and reconstruction or "assembly" of whole genome sequences. Accurate genome assembly in model organisms (organisms that are amenable to genetic study, have short generation times, breed in large numbers, and can inform about other organisms) has traditionally relied upon an edifice of classical genetics resources including inbred lines to minimize genetic variation, genetic linkage maps generated from laboratory crosses among inbred lines, and the sequencing and hierarchical or clone-based assembly of large 40–200 kb genome fragments called "bacterial artificial chromosomes" (BACs) large-insert libraries (Lander et al. 2001). Arachnids, like most non-model organisms, lack most of these resources. They often have long generation times and can be very small (forcing pooling of individuals and an increase in heterozygosity, making assembly difficult). Moreover, they are generally difficult to breed in captivity and, except for some mite species, no inbred lines are available, with the possible exceptions of naturally inbred social species such as the eresid *Stegodyphus mimosarum* (J.S. Bechsgaard & T. Bilde pers. comm.) and theridiid *Anelosimus eximius* (I. Agnarsson pers. comm.).

5.1. Published genomes.—The three presently available arachnid genomes are from highly derived acarine species: the two-spotted spider mite *Tetranychus urticae* (Grbić et al. 2011), the honey bee ectoparasitic mite *Varroa destructor* (Cornman et al. 2010) and the deer tick *Ixodes scapularis* (http://iscapularis.vectorbase.org). The choice of these arachnids as early targets for genome sequencing is perhaps unsurprising; *Tetranychus* and *Varroa* are of tremendous agricultural and economic importance, and *Ixodes* is of great importance as a vector of numerous livestock and human diseases including Lyme disease. In addition to its economic importance, *Tetranychus urticae* was selected as a candidate for genome sequencing as it has the smallest known genome of any arthropod at a mere 89.6 Mbp (Grbić et al. 2011), is easily cultured in the laboratory and has inbred lines available. The small *Tetranychus* genome was sequenced using traditional Sanger sequencing methods to a depth of 8.05X, resulting in 640 scaffolds: 70,778 EST sequences plus RNA-seq data (see below) were mapped to the genome and supported 15,397 of 18,414 gene models. The genome of the ectoparasitic mite *Varroa destructor*, which has emerged as the primary pest of domestic honey bees (Cornman et al. 2010), was "surveyed" using 4.3X coverage of 454 sequence data from the DNA of 1,000 pooled mites. This 2.4 Gbp was clearly insuffieient to provide a comprehensive de novo assembly of this moderately sized genome (at 294 Mbp still far bigger than most sequenced insects) and yielded 184,094 contigs (assembled contiguous but not "scaffolded" sequences) with an N50 (weighted median of contig lengths) of 2,626 bp; however, the data were sufficient to permit the prediction of 31.3 Mbp of gene sequence, information about the integration of microbes into the genome and the occurrence of single nucleotide polymorphisms (Cornman et al. 2010). Finally, the genome of the deer tick *Ixodes scapularis*, which is very large compared to *Tetranychus* and *Varroa* at 2.1 Gbp, was shotgun sequenced using Sanger sequencing to a coverage of 3–6X. Although many data on the expressed gene sequences (i.e., the transcriptome) are available in the public databases, the genome sequence remains highly fragmented (e.g., ca. 571,000 contigs with a contig N50 of 3000 bp) and has not been officially published (http://iscapularis.vectorbase.org). Of the three acarine species, *Tetranychus* provides the most complete genome reconstruction, with genome assemblies for *Varroa* and *Ixodes* remaining highly fragmented.

5.2. Genomes in progress – what have we learned?—Apart from the three acarine species discussed above, our knowledge of the nuclear DNA structure of arachnids remains extremely limited. Most knowledge about arthropods comes from insects—a reflection of biological diversity, societal impact and economic and medical importance, and the scale of the research community, among other factors. From an evolutionary and phylogenetic perspective, this bias of course does not reflect relative importance. However, efforts such as the research community-driven I5K project (http://www.arthropodgenomes.org/wiki/Main_Page) that aims to sequence 5,000 arthropod genomes over five years should redress the balance to some extent. Even so, of 787 species currently nominated for sequencing, there are 702 Hexapoda (89%), 64 Chelicerata (8%), only 20 Crustacea (2%), and 6 Myriapoda (1%) (http://www.arthropodgenomes.org/wiki/i5K_nominations). Several arachnids have been included in the pilot sequencing project of the I5K, and these are discussed below, together with our own efforts on *Theridion* (Theridiidae) and other efforts on *Stegodyphus* (Eresidae) and *Acanthoscurria* (Theraphosidae). In addition, the genome of *Limulus* has been sequenced, and a preliminary assembly is about to be publicly released (Nipam Patel pers. comm.). However, pre-NGS we have revealed much about arachnid mitochondrial genomes, and we briefly review this here before going on to examine nuclear genomes.

Mitochondrial genomes: Although knowledge of the arachnid nuclear genome remains in its infancy, several decades of research, based upon traditional PCR and Sanger sequencing, have yielded detailed knowledge of arachnid mitochondrial genomes. This work has revealed lineage-specific gene order rearrangements in Opiliones (Masta 2010) and pseudoscorpions (Ovchinnikov & Masta 2012), and most interestingly has revealed truncated mitochondrial tRNA (and rRNA) secondary structures among most arachnid lineages (Masta & Boore 2004, 2008; Masta et al. 2008; Fahrein et al 2009; Masta 2010; Ovchinnikov & Masta 2012). NGS technologies can potentially greatly increase our understanding of the sequence diversity, variation and transcriptional mechanisms among arachnid mitochondria since 1) whole mitochondrial genomes can rapidly be sequenced from many barcoded and pooled individuals using amplicon sequencing (e.g., on small scale MiSeq or IonTorrent systems) (see below); 2) mitochondrial genomes can be assembled from total genome sequence data (Iorizzo et al. 2012); and 3) RNA-seq reads (Illumina-based method of sequencing cDNA obtained via reverse transcription of mRNA extractions) can be mapped to mitochondrial genes to explore expression differences among genes and taxa and post-transcriptional modification and editing of gene sequences (Smith 2013).

Nuclear genomes: Since no non-acarine genomes have been published so far, detailed discussion of their structure is not yet possible. Our own efforts at sequencing a spider genome have focused on the Hawaiian happy face spider *Theridion grallator* and primarily have used Illumina paired end data based upon a variety of insert sizes. Initial assemblies were highly fragmented (resulting in many contigs of short length; i.e., a low "contig N50"). Although this is partly due to heterozygosity (no inbred lines are available), the main complication appears to be that this species has a low average % GC across the genome (ca. 28%) (Fig. 1). Although most arthropod genomes are somewhat "AT-rich" (e.g., the honey bee *Apis mellifera* has 34.8% GC; The Honeybee Genome Sequencing Consortium 2006), the only arthropod genome with a % GC in the range we have found is that of the pea aphid *Acyrthosiphum pisum* at 29.6% GC (The International Aphid Genomics Consortium 2010).

A potential extreme % GC bias in arachnid genomes is both intriguing and technically challenging from both an informatic and molecular biological point of view. In order to investigate this further, we have examined the assembled contigs data, where available, from the pilot runs for the I5K project (http://www.arthropodgenomes.org/wiki/Main_Page). In total we have examined the contig N50 length and % GC from two I5K sequenced spiders, *Latrodectus hesperus* and *Parasteatoda tepidariorum* (Theridiidae), and the Arizona bark scorpion *Centruroides sculturatus* (Buthidae), together with 15 other arthropods (14 insects and one copepod; ftp://ftp.hgsc.bcm.edu/I5K-pilot/), and these are plotted in Fig. 1. In addition we have included our data from *Theridion grallator* (PJPC unpubl. data) and data from *Stegodyphus mimosarum* (Eresidae: J.S. Bechsgaard & T. Bilde, pers. comm.). The scorpion and the three theridiid spiders (*L. hesperus*, *P. tepidariorum* and *T. grallator*) all have less than 30 % GC and correspondingly low contig N50 lengths. In general, the lower the % GC, the shorter the contig N50 as a simple function of decreased information content available to the assembly algorithms. Interestingly, the *P. tepidariorum* sequenced had been through five generations of inbreeding (A. McGregor pers. comm.)—apparently sufficient to reduce heterozygosity enough to substantially increase contig lengths despite a low % GC.

Alternatively, *S. mimosarum* did not exhibit an extreme % GC bias (34% GC) and as a soeial species (Mattila et al. 2012) is somewhat naturally inbred—and has a correspondingly

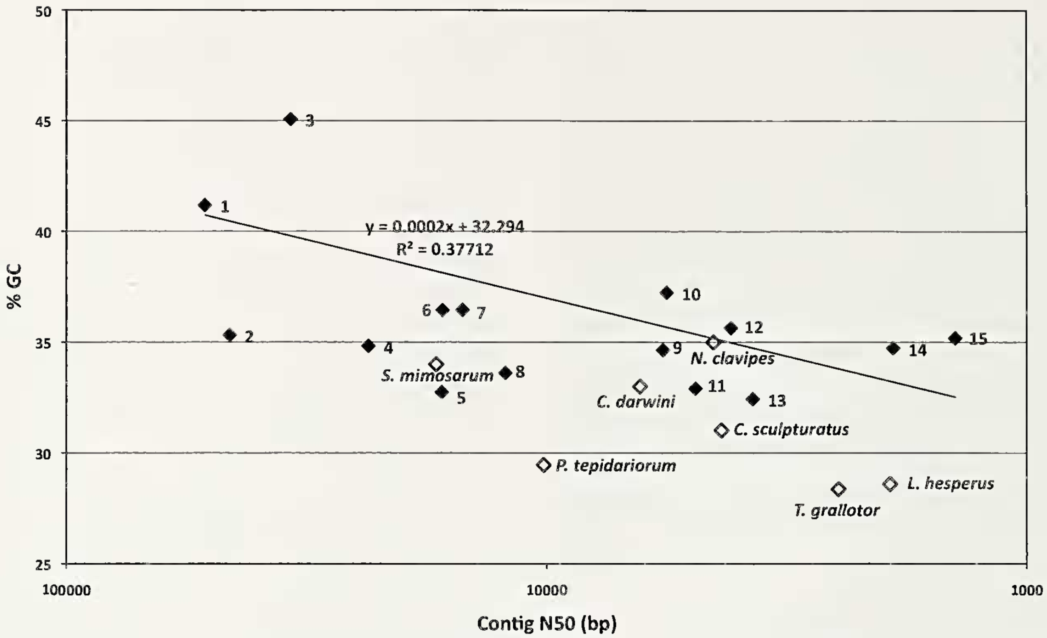

Figure 1.—Assembly contig N50 length (bp) is negatively correlated with average genome-wide %GC bias among arthropod taxa. As average %GC decreases so does the contig N50 (weighted median of contig lengths), since lower information content leads to more fragmented assemblies. Open diamonds refer to arachnid genome projects and closed diamonds refer to other arthropods. *Theridion grallator* data are calculated from the authors' own (unpublished) preliminary genome assembly and have the lowest %GC to date (28.37%). *Stegodyphus mimosarum* values from J.S. Bechsgaard (pers. comm.). *C. darwini* and *N. clavipes* values from I. Agnarsson (pers. comm.). All other values estimated from initial contig (not scaffolded) assemblies of the I5K initiative pilot genome assemblies; data and assembly parameters are therefore similar among species (ftp.hgsc.bcm.edu/I5K-pilot/). Although all arthropods show a bias toward low %GC (<32%), the theridiid spiders *T. grallator*, *L. hesperus* and *P. tepidariorum*, as well as the scorpion *C. sculpturatus*, have very low %GC. *S. mimosarum* has a more moderate bias (34% GC), and both this species and *P. tepidariorum* show the benefit of inbreeding and low-heterozygosity and have longer contig N50 lengths than the other arachnids. Additionally, the remaining non-theridiid spiders, *C. darwini* and *N. clavipes*, have moderate %GC values but low contig N50s (<10,000 bases), possibly due to heterozygosity stemming from the lack of inbreeding. The included insects are 1) *Athalia rosae* (turnip sawfly: Insecta: Hymenoptera); 2) *Ceratitis capitata* [Mediterranean fruitfly (medfly): Insecta: Diptera]; 3) *Orussus abietinus* (parasitic wood wasp: Insecta: Hymenoptera); 4) *Cimex lectularius* (bed bug: Insecta: Hemiptera); 5) *Anoplophora glabripennis* (Asian long-horned beetle: Insecta: Coleoptera); 6) *Libellula fulva* (scarce chaser: Insecta: Odonata); 7) *Helicoverpa punctigera* (Australian bollworm: Insecta: Lepidoptera); 8) *Ephemera danica* (green drake mayfly: Insecta: Ephemeroptera); 9) *Agrilus planipennis* (emerald ash borer: Insecta, Coleoptera); 10) *Copidosoma floridanum* (chalcid wasp: Insecta: Hymenoptera); 11) *Homalodisca vitripennis* (glassy-winged sharpshooter: Insecta: Hemiptera); 12) *Leptinotarsa decemlineata* (Colorado potato beetle: Insecta: Coleoptera); 13) *Eurytemora affinis* [copepod: Maxillopoda (Crustacea): Calanoida]; 14) *Limnephilus lunatus* (caddisfly: Insecta: Trichoptera); 15) *Pachypsylla venusta* (hackberry petiole gall psyllid: Insecta: Hemiptera).

much better assembly contiguity (J.S. Bechsgaard & T. Bilde pers. comm.). Furthermore, initial sequencing of the huge (6 Gbp) genome of the Brazilian white knee tarantula *Acanthoscurria geniculata* (Theraphosidae) indicates that this species has a ca. 40% GC genome content (J.S. Bechsgaard & T. Bilde pers. comm.). Until more arachnid genome sequence becomes available, the question as to how widespread %GC-bias is among arachnids will remain unclear. From the above data it may appear to be specific to theridiid spiders and *Centruroides* scorpions; however it is tempting to speculate that extreme %GC-bias may extend to other spider families and other arachnid orders. This possibility should be considered in future genome-sequencing efforts, and we note that transcriptome assembly (RNA-seq) is unlikely to be so impacted by %GC-bias, since coding regions typically do not exhibit such extreme biases.

5.3. A cautionary note.—Despite the allure of NGS technologies, some caution is needed before embarking on a project to sequence an arachnid genome. Particular questions a researcher working on a specific taxon should pose are: 1) Do we need a genome sequence? And if so, 2) how complete do we need it to be? And, perhaps more fundamentally, 3) what level of completeness can we attain without spending an unreasonable amount of resources? In reality, no genome sequence (including human) is fully complete, and de novo assembled and NGS derived genomes are even less so. De novo assembly of short-read shotgun sequence data without references, such as linkage maps or BAC libraries, remains extremely challenging. However, as the *Tetranychus*, *Varroa*, and *Ixodes* projects demonstrate, a fractured assembly may still be useful if it is contiguous enough to build valid gene models. In addition to life history and often body-size considerations (i.e., the need for pooling individuals), intrinsic features of arachnid genomes—in particular, the low % GC content in some lineages mentioned above—might raise a substantial barrier to whole genome de novo assembly projects.

Even though the cost of NGS sequencing continues to drop rapidly, depending upon the biological question, either classical genetic marker-based approaches (Section 3 above) may be cheaper, easier, and sufficient, or NGS based alternatives to genome sequencing may be more attainable (e.g., transcriptome sequencing, and reduced representation methods). Indeed these approaches may even be best used as a means to rapidly develop numerous classical markers or identify single nucleotide polymorphisms (as discussed in Section 6). RNA-seq (the sequencing of cDNA libraries derived from extracted mRNA and hence targeting transcribed and therefore mainly coding regions—the transcriptome) is rapidly becoming the tool of choice in genomic studies. This is because RNA-seq data permits one both to build gene models rapidly and to measure "digital" gene expression among taxa and tissues; consequently the technique has many potential applications.

Although "complete" genome sequences, even fragmented ones, will yield fascinating information about genome structure (repeats, transposons, translocations, etc.), to be of greatest functional utility genomes must be annotated. While computational annotation of gene models is possible (although of course not optimized for arachnids), most annotation schemes work best when supported by sequence evidence. Again, RNA-seq and transcriptome data are of greatest utility here and thus should also be generated for the taxon whose genome is sequenced. Since RNA-seq data can be assembled de novo, for example using software Trinity (Grabherr et al. 2011), and annotated by homology searches (at least for genes where known homologs exist) (e.g., using BlastX and Blast2GO; Conesa et al. 2005), the experimenter must again ask whether a full genome sequence is required at all, and be cautious about assuming that this is a practicable route.

6. APPLICATIONS OF NGS TECHNOLOGIES IN ARACHNOLOGY

The number of possible applications using NGS technologies is vast and continues to grow. Here we provide examples of their use, most of which do not require the sequencing of entire genomes. There are many more potential applications than those discussed below, and, as new platforms and bioinformatic tools are developed, new avenues of research will open.

6.1. Functional genomics: adaptation & selection.—Biologists frequently seek to elucidate the relationship between environmental parameters and organismal diversity. The potential for detailing the genetic response of an organism to changes in the biotic and abiotic environment are now in plain sight with the availability of vast quantities of DNA sequence that can be generated by NGS technologies, in particular through the "assembly" of whole genome sequences. A review by Stapley et al. (2010) discusses the potential of high-throughput technologies in studies of adaptation. Whether focusing on coding gene sequences, differential expression of transcripts, identifying genomic regions experiencing linkage disequilibrium (LD), or quantitative trait locus (QTL) mapping to detect genomic regions under selection, established methods using high-throughput data exist.

Measures of selection: When studying protein-coding loci, the most common method for measuring selection involves the ratio of nonsynonymous to synonymous changes (dN/dS or ω). The resulting value potentially indicates the mode of selection acting on the gene: $\omega = 1$ (neutral selection), $\omega < 1$ (stabilizing selection) and $\omega > 1$ (positive selection). By employing a likelihood ratio test, P-values can be obtained to differentiate between neutral and directional selection in pairwise comparisons. Additionally, comparisons of ω between branches in a multi-species/population phylogeny are possible to identify genes or residues evolving differently or similarly between clades. Inherently, ω-based tests for selection require coding data and are best served by transcriptomic data. Commonly used tools for analysis of these data include PAML (Yang 2007) and HyPhy (Pond et al. 2005). Some studies have employed ω tests in arthropods (e.g., Averof 2002; Porter et al. 2006; Viljakainen et al. 2009; Fort et al. 2011), recently including spiders (Brewer et al. in review; Yim et al. in prep).

To collect the data necessary for investigating selection in coding sequences, RNAseq libraries are often generated. To obtain the most nearly unique sequence possible in a single run, the resulting cDNA libraries can be normalized by removing excessive copies of highly expressed transcripts to "equalize" the numbers with respect to the more poorly-expressed transcripts (Zhulidov et al. 2005), but normalizing is not essential. In addition to retrieving sequences, non-normalized RNAseq libraries provide information concerning the expression levels of transcripts. In order to leverage this information, specimens must be treated to control all variables so that the sources of differential expression (DE) can be identified. Methods to analyze expression data using RNA-seq data include the R packages "edgeR" (Robinson et al. 2009) and "DEseq" (Anders & Huber 2010). Differences in expression of transcripts between populations or species indicate the evolution of coding loci involved in the expression of a gene or non-coding regions of the genome that affect the transcription (i.e., promoters, enhancers, and suppressors). These methods are currently being employed in Hawaiian *Tetragnatha* spiders to study differences in venom composition between a lineage that builds webs compared to one that does not build webs (Brewer et al. in review), building on earlier work that used protein gel electrophoresis patterns to show coarse differences between these lineages (Binford 2001). With NGS techniques, we are now able to explore the individual genes and relative changes in expression levels.

Selection can also be examined using LD approaches, although this is necessarily limited to taxa where full genomes are available. By mapping SNPs to a reference genome, data obtained using reduced representation techniques (e.g., RAD-seq) can be used to detect regions of the genome under strong LD. This method has been used to identify regions of the genomes of stickleback populations that are resistant to introgression of outside genes (Hohenlohe et al. 2010). Unfortunately, RAD-seq methods require high quality and high quantity gDNA, which is often limited in small organisms such as many spiders, even when freshly collected (Cotoras unpubl. data).

Molecular basis for adaptation: Perhaps the most important applications of NGS technologies in arachnids relate to silk and venoms, two aspects of the biology of these organisms that provide an almost endless variety of questions relating to gene function. Both silks and venoms comprise complex combinations of highly-derived, and often highly repetitive, proteins that serve myriad functions within and between taxonomic groups. Both are linked to major ecological shifts and evolutionary modifications in a number of clades. The evolution of the forms and functions of spider silks has great potential in evolutionary studies, as well as bioengineering applications (Blackledge 2012; Garb 2013). Tools such as SMRT and Nanopore, with their long reads, could help to alleviate assembly issues associated with the highly repetitive elements and allow more detailed exploration of the diversity of spider silks at the genomic level. Venoms also vary greatly across the Arachnida and are found in several orders (e.g., Araneae, Scorpiones, and Pseudoscorpiones). Beyond differential expression analyses, such as that described above, characterization of venom cocktails and their molecular evolution is lacking in most groups. Most work done so far has focused on medically relevant species such as those in the spider genera *Latrodectus* (e.g., Garb & Hayashi 2013) and *Loxoceles* (e.g., Zobel-Thropp et al. 2013) and the scorpion genus *Centruroides* (e.g., Valdez-Velázquez et al. 2013). In an applied context, these compounds have vast potential in pharmacology and as pest control substances (reviewed by King and Hardy 2013). Moreover, as mentioned above, these compounds may also provide insights into the factors underlying adaptation and how selection acts at the transcriptional level (Binford 2001).

6.2. Phylogenetics.—To date, most molecular studies of arachnids have sought to ascertain relationships between taxa. Until recently, assessment of the phylogenetic affinities of an organism required PCR amplification with degenerate primers followed by amplicon sequencing to study loci in distantly related taxa. The weakness of this approach is that rather few loci can be examined, limiting the resolution of the Tree of Life. Thus, the internal phylogeny of the subphylum Chelicerata, class Arachnida, and lower taxonomic levels has remained unresolved despite numerous efforts to ascertain the relationships between taxa, including molecular phylogenetic studies (recently reviewed in Agnarsson et al. 2013; Giribet and Edgecombe 2013). Mitochondrial sequences have been the most common data source. However, for the reasons mentioned above (3.1), mitogenomic sequence data may not be appropriate for reconstructing deep arthropod relationships (Brewer et al. 2013). For example, although the Euchelicerata (Xiphosura + Chelicerata) is almost unambiguously recovered using nuclear loci, datasets using mitochondrial genomic data often fail to support this relationship (Masta et al. 2009; Rota-Stabelli 2010).

Within the Arachnida, most molecular phylogenetic studies have focused on spiders, including the relationships within the subclasses Mygalomorphae (Hedin and Bond 2006; Bond et al., 2012) and Araneomorphae (Blackledge et al. 2009). Molecular phylogenetic analyses within other orders exist, including the Opiliones (Hedin et al. 2010; Hedin et al. 2012; Burns et al. 2013), Acari (Klompen et al. 2007; Dabert et al. 2010; Pepato et al. 2010), Scorpiones (Salomone et al. 2007; Borges et al. 2010; Prendini & Esposito 2010) and Amblypygi (Esposito et al. in review). Representing a small sampling of published works, all of these studies except Hedin et al. (2012) use traditional Sanger sequencing approaches. Even at these lower taxonomic levels, nuclear molecular markers with appropriate phylogenetic signal are lacking, and primer combinations for PCR often do not transfer between arachnid groups, especially for species/population-level appropriate loci.

High-throughput sequencing technologies provide a means to collect vast amounts of molecular data for many taxa in a timely manner and are currently used in various ways in phylogenetics (see McCormack et al. 2013 and Rocha et al. 2013). The potential use of some NGS technologies in spider systematics was recently discussed by Agnarsson et al. (2013) and in Opiliones by Hedin et al. (2012). As for most non-model organisms, the most common NGS data sources for deep phylogenetics in arachnids are transcriptomes (Agnarsson et al. 2013) and information generated from bait capture techniques (for all taxonomic levels) such as anchored enrichment (Lemmon et al. 2012). These approaches do not require full genome sequences, which is especially useful given the potential difficulties with arachnid genome efforts mentioned above; moreover, the data generated provide loci that are relatively easy to assign orthology and can be used at deep taxonomic levels. Tools for the assignment of orthology include HaMStR (Ebersberger et al. 2009), OrthoDB (Waterhouse et al. 2012) and AGALMA (Dunn et al. 2013), while PhyDesign (López-Giráldez and Townsend 2011) can be used to investigate the phylogenetic signal of a locus across an ultrametric tree. Recent molecular models of evolution (e.g., CAT, Lartillot and Phillippe 2004) and algorithms for phylogeny reconstruction (e.g., Phylobayes, Lartillot et al. 2009; RAxML, Stamatakis 2006; and Fasttree 2, Price et al. 2010) have made phylogenomic studies much more tractable. However, these analyses still can take weeks of computation time, require large amounts of computer memory, and demand a somewhat deep understanding of bioinformatics.

6.3. Population genetics & phylogeography.—NGS approaches have been widely celebrated for their potential in providing large numbers of markers across the genome, which is essential for population genetic and phylogeographic studies. Since the per base cost is much lower than for Sanger sequencing, it has become economical to apply NGS techniques to generate traditional markers [e.g., microsatellites in a tetragnathid species (Parmakelis et al. 2013)].

Among the most useful tools for population genetics, and phylogenetics for that matter, are those based on reduced representation libraries (RRLs), which attempt to recover a small, random (i.e., unlinked) snapshot of the total genome. As a result of focusing on a small sample of the genome, the cost of sequencing a single individual is greatly reduced and yet RRL methods can still identify many thousands of usable single nucleotide polymorphisms (SNPs).

RADseq is a popular method for genome-wide marker analysis because it reduces the complexity of the genome by sub-sampling at certain restriction sites, assumed to be homologous among taxa/specimens, to generate a single nucleotide polymorphism (SNP) data set. The approach is much like RFLPs and AFLPs, except that, instead of separating the fragments on a gel to recover a DNA fingerprint, they are sequenced (Davey et al. 2011). This approach can provide several SNPs from each fragment, multiplying the amount of data obtained from a single run. A recent modification of this technique uses a double-digestion and yields an increase in efficiency and a reduction in cost

(Peterson et al. 2012). However, for many arachnid groups, the RADseq method has requirements that may limit its use. First, a large amount of high molecular weight DNA is required (>2 micrograms per sample). Such high quality DNA is essential in order to generate fragments that result only from the restriction enzyme digest (i.e., library adaptors are not ligated to the end of randomly sheared/degraded fragments). Moreover, a high starting concentration of DNA is necessary, since the protocol involves many steps that result in the loss of DNA—typically only 7–15% of the starting material will be recovered. The issue of DNA quality can be resolved by preserving samples in 95% ethanol at -80 C, RNAlater at −80 C, or by using fresh specimens. Standard extraction kits using ion-exchange columns or salt precipitation should work well without causing undue shearing of the DNA. Ultimately, DNA yield depends on the organism and although large-bodied arachnids (e.g., many mygalomorphs, scorpions, amblypigids, or solfugids) may yield sufficient DNA, smaller taxa may require specimens to be "pooled" together, thus losing individual-level data (Emerson et al. 2010).

An alternative RRL approach to RADseq is to use bait capture methods, such as Exon Capture, which, in contrast to RADseq, requires starting material to be randomly fragmented (Bi et al., 2012). The basic approach is to sequence the transcriptome of one individual and use those sequences to design small, overlapping probes that are then attached to a capture array ("chip") or beads. The protocol starts with either naturally (i.e., degraded or historical) or intentionally fragmented DNA, which is used to prepare DNA libraries following a standard NGS protocol (e.g., Meyer & Kircher 2010). These libraries are barcoded for each individual and used in a hybridization experiment similar to a microarray. The number of individuals that can be multiplexed in these experiments depends on the target size in base pairs (i.e., the number of bases printed on the chip), the desired depth of coverage and the availability of barcodes. The number of single barcodes commercially available is currently 96, but by using eight double barcodes this can be increased to 768. This theoretically allows the parallel sequencing of thousands of loci in hundreds of individuals at the same time. One advantage of exon capture for non-model organisms is that the sequences for the array are obtained directly from a transeriptome and do not require a previously sequenced genome. Moreover, the protocol is suitable for historical museum samples, since it explicitly requires randomly fragmented DNA, which is often the natural state for museum-derived material.

There are two main limitations to the Exon Capture approach. The first is that starting costs are high (reagents and specialized equipment not common in most laboratories), though they can be minimized by sharing among research groups. Second, as in most NGS applications, sophisticated expertise in bioinformatics is needed to manage the large and complex data sets. Fortunately, user-friendly programs and tools are becoming increasingly available for post-NGS processing and analyses. Since exon capture targets exons, most of the captured variation will correspond to synonymous mutations in coding genes, allowing insights into population variability. However, because genomic DNA is captured, some of the non-coding flanking regions (e.g., untranslated regions, introns) will also be recovered.

7. CONCLUSION

Arachnids have a rich history of molecular studies focusing on many aspects of their biology. To date, few of these have made use of recent advances in sequencing technology, but, as we have outlined above, many future projects should benefit from the use of next-generation sequencing platforms. These technologies are diverse in their methods and applications, and promising advances are on the horizon. However, it is important to realize the strengths and weakness of NGS tools and to embrace traditional techniques when more appropriate.

Although it is easy to be seduced by the amount of data that can be generated by sequencing an entire genome, this is often not necessary. In many cases, studies using transcriptomes or reduced representation techniques can collect incredible amounts of useful data to address any number of questions. Regardless of the study, the number of potential avenues to gather molecular data is large in terms of strategy and scale. As arachnologists continue to amass novel data from diverse lineages, our ability to identify loci, in terms of function and homology, will increase and open more research opportunities. The unique biology and evolutionary history of arachnids, coupled with technological and bioinformatic advances, will provide research opportunities for years to come.

ACKNOWLEDGMENTS

We would like to thank Douglass Morse for shepherding this publication, two reviewers (Marshal Hedin and Ingi Agnarsson) for greatly enhancing this work, and the I5k initiative (J.S. Bechsgaard, J. Bechsgaard, A. McGregor and T. Bilde) for providing preliminary data used herein. D. Cotoras is funded by a Fulbright/CONICYT scholarship. The *Theridion grallator* genome project is funded by NSF DEB 0919215.

LITERATURE CITED

A'Hara, S., R. Harling, R. McKinlay & C. Topping. 1998. RAPD profiling of spider (Araneae) DNA. Journal of Arachnology 26:397–400.

Agnarsson, I., J.A. Coddington & M. Kuntner. 2013. Systematics: Progress in the study of spider diversity and evolution. Pp. 58–111. *In* Spider Research in the 21st Century: Trends and Perspectives. (D. Penney, ed.). Siri Scientific Press, Manchester, UK.

Amaral-Zettler, L.A., E.A. McCliment, H.W. Ducklow & S.M. Huse. 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS ONE 4:e6372.

Anders, S. & W. Huber. 2010. Differential expression analysis for sequence count data. Genome Biology 11:R106.

Averof, M. 2002. Arthropod Hox genes: insights on the evolutionary forces that shape gene functions. Current Opinion in Genetics & Development 12:386–392.

Baert, L., F. Hendrickx & J.P. Maelfait. 2008. Allozyme characterization of *Hogna* species (Araneae, Lycosidae) of the Galapagos Archipelago. Journal of Arachnology 36:411–417.

Bi, K., D. Vanderpool, S. Singhal, T. Linderoth, C. Moritz & J.M. Good. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics 13:403.

Bilde, T., C. Tuni, A. Cariani, A. Santini, C. Tabarroni & F. Garoia, et al. 2009. Characterization of microsatellite loci in the subsocial spider *Stegodyphus lineatus* (Araneae: Eresidae). Molecular Ecology Resources 9:128–130.

Bilde, T., Y. Lubin, D. Smith, J.M. Schneider & A.A. Maklakov. 2005. The transition to social inbred mating systems in spiders: role of inbreeding tolerance in a subsocial predecessor. Evolution 59:160–174.

Binford, G. 2001. Differences in venom composition between orb-weaving and wandering Hawaiian *Tetragnatha* (Araneae). Biological Journal of the Linnean Society 74:581–595.

Blackledge, T.A. 2012. Spider silk: a brief review and prospectus on research linking biomechanics and ecology in draglines and orb webs. Journal of Arachnology 40:1–12.

Blackledge, T.A., N. Scharff, J.A. Coddington, T. Szuts, J.W. Wenzel & C.Y. Hayashi, et al. 2009. Reconstructing web evolution and spider diversification in the molecular era. Proceedings of the National Academy of Sciences of the United States of America 106:5229–5234.

Blick, T. & M.S. Harvey. 2011. Worldwide catalogues and species numbers of the arachnid orders (Arachnida). Arachnologische Mitteilungen 41:41–43.

Bond, J.E., B.E. Hendrixson, C.A. Hamilton & M. Hedin. 2012. A reconsideration of the classification of the spider infraorder Mygalomorphae (Arachnida: Araneae) based on three nuclear genes and morphology. PLoS ONE 7:e38753.

Borges, A., E. Bermingham, N. Herrera, M.J. Alfonzo & O.I. Sanjur. 2010. Molecular systematics of the neotropical scorpion genus *Tityus* (Buthidae): The historical biogeography and venom antigenic diversity of toxic Venezuelan species. Toxicon 55:436–454.

Boulton, A.M., M.G. Ramirez & C.P. Blair. 1998. Genetic structure in a coastal dune spider (*Geolycosa pikei*) on Long Island, New York Barrier Islands. Biological Journal of the Linnean Society 64:69–82.

Brewer, M.S., L. Swafford, C.L. Spruill & J.E. Bond. 2013. Arthropod phylogenetics in light of three novel millipede (Myriapoda: Diplopoda) mitochondrial genomes with comments on the appropriateness of mitochondrial genome sequence data for inferring deep level relationships. PLoS ONE 8:e68005.

Burns, M.M., M. Hedin & J.W. Shultz. 2013. Comparative analyses of reproductive structures in harvestmen (Opiliones) reveal multiple transitions from courtship to precopulatory antagonism. PLoS ONE 8:e66767.

Caporaso, J.G., C.L. Lauber, W.A. Walters, D. Berg-Lyons, J. Huntley & N. Fierer, et al. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME Journal 6:1621–1624.

Conesa, A., S. Götz, J.M. Garcia-Gomez, J. Terol, M. Talon & M. Robles. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676.

Cornman, R.S., M.C. Schatz, J.S. Johnston, Y-P. Chen, J. Pettis & G. Hunt, et al. 2010. Genomic survey of the ectoparasitic mite *Varroa destructor*, a major pest of the honey bee *Apis mellifera*. BMC Genomics 11:602.

Croucher, P.J.P., G.S. Oxford & R.G. Gillespie. 2011a. Population structure and dispersal in a patchy landscape: nuclear and mitochondrial markers reveal area effects in the spider *Theridion californicum* (Araneae: Theridiidae). Biological Journal of the Linnean Society 104:600–620.

Croucher, P.J.P., G.S. Oxford, A. Lam & R.G. Gillespie. 2011b. Stabilizing selection maintains exuberant colour polymorphism in the spider *Theridion californicum* (Araneae, Theridiidae). Molecular Ecology 20:206–218.

Croucher, P.J.P., G.S. Oxford, A. Lam, N. Mody & R.G. Gillespie. 2012. Colonization history and population genetics of the color-polymorphic Hawaiian happy-face spider *Theridion grallator* (Araneae, Theridiidae). Evolution 66:2815–2833.

Dabert, M., W. Witalinski, A. Kazmierski, Z. Olszanowski & J. Dabert. 2010. Molecular phylogeny of acariform mites (Acari, Arachnida): strong conflict between phylogenetic signal and long-branch attraction artifacts. Molecular Phylogenetics and Evolution 56:222–241.

Dahl, F., J. Stenberg, S. Fredriksson, K. Welch, M. Zhang & M. Nilsson, et al. 2007. Multigene amplification and massively parallel sequeneing for cancer mutation discovery. Proceedings of the National Academy of Sciences of the United States of America 104:9387–9392.

Davey, J.W., P.A. Hohenlohe, P.D. Etter, J.Q. Boone, J.M. Catchen & M.L. Blaxter. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12:499–510.

De Busschere, C., F. Hendrickx, S.M. Van Belleghem, T. Backeljau, L. Lens & L. Baert. 2010. Parallel habitat specialization within the wolf spider genus *Hogna* from the Galapagos. Molecular Ecology 19:4029–4045.

Duncan, S.I., S.E. Riechert, B.M. Fitzpatrick & J.A. Fordyce. 2010. Relatedness and genetic structure in a socially polymorphic population of the spider *Anelosimus studiosus*. Molecular Eeology 19:810–818.

Dunn, C.W., M. Howison & F. Zapata. 2013. Agalma: an automated phylogenetics workflow. BMC Bioinformatics 141:330.

Ebersberger, I., S. Strauss & A. von Haeseler. 2009. HaMStR: Profile hidden Markov model based search for orthologs in ESTs. BMC Evolutionary Biology 9:157.

Emerson, K.J., C.R. Merz, J.M. Catchen, P.A. Hohenlohe, W.A. Cresko & W.E. Bradshaw, et al. 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proceedings of the National Academy of Sciences of the United States of America 107:16196–16200.

English, A.C., S. Richards, Y. Han, M. Wang, V. Vee & J. Qu. et al. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS ONE 7:e47768.

Evans, T.A. & M.A.D. Goodisman. 2002. Nestmate relatedness and population genetic structure of the Australian social crab spider *Diaea ergandros* (Araneae: Thomisidae). Molecular Ecology 11:2307–2316.

Fahrein, K., S.E. Masta & L. Podsiadlowski. 2009. The first complete mitochondrial genome sequences of Amblypygi (Chelicerata: Arachnida) reveal a conservation of the ancestral arthropod gene order. Genome 52:456–466.

Fort, P., A. Albertini, A. Van-Hua, A. Berthomieu, S. Roehe & F. Delsuc, et al. 2011. Fossil rhabdoviral sequences integrated into arthropod genomes: ontogeny, evolution, and potential functionality. Molecular Biology and Evolution 29:381–390.

Garb, J.E. 2013. Spider Silk: An ancient biomaterial for the 21st century. Pp. 252–281. *In* Spider Research in the 21st Century: Trends and Perspectives. (D. Penney, ed.). Siri Scientific Press, Manchester, UK.

Garb, J.E. & C.Y. Hayashi. 2013. Molecular evolution of latrotoxin, the exceptionally potent vertebrate neurotoxin in black widow spider venom. Molecular Biology and Evolution 30:999–1014.

Garb, J.E. & R.G. Gillespie. 2009. Diversity despite dispersal: colonization history and phylogeography of Hawaiian crab spiders inferred from multilocus genetic data. Molecular Ecology 18:1746–1764.

Gillespie, R.G., H.B. Croom & S.R. Palumbi. 1994. Multiple origins of a spider radiation in Hawaii. Proceedings of the National Academy of Sciences of the United States of America 91:2290–2294.

Gillespie, R.G. 2013. Biogeography: From testing patterns to understanding processes in spiders and related arachnids. Pp. 154–185. *In* Spider Research in the 21st Century: Trends and Perspectives. (D. Penney, ed.). Siri Scientific Press, Manchester, UK.

Giribet, G. & G.D. Edgecombe. 2013. The Arthropoda: a phylogenetic framework in arthropod biology and evolution. Pp. 17–40. *In*

Arthropod Biology and Evolution. (A. Minelli, ed.). Springer-Verlag, Berlin.

Gloor, G.B., R. Hummelen, J.M. Macklaim, R.J. Dickson, A.D. Fernandes & R. MacPhee, et al. 2010. Microbiome profiling by Illumina sequencing of eombinatorial sequence-tagged PCR products. PLoS ONE 5:e15406.

Goodacre, S.L. 2013. Genetics and Genomics: The arrival of a new molecular era of spider research. Pp. 186–199. *In* Spider research in the 21st century: trends and perspectives. (D. Penney, ed.). Siri Scientific Press, Manchester, UK.

Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson & I. Amit, et al. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nature Biotechnology 29:644–652.

Grbić, M., T. Van Leeuwen, R.M. Clark, S. Rombauts, P. Rouzé & V. Grbić, et al. 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature 479:487–492.

Gurdebeke, S., B. Neirynck & J.P. Maelfait. 2000. Population genetic effects of forest fragmentation in Flanders (Belgium) on *Coelotes terrestris* (Wider) (Araneae: Agelenidae) as revealed by allozymes and RAPD. Ekologia-Bratislava 19:87–96.

Gurdebeke, S., J.P. Maelfait & T. Backeljau. 2003. Contrasting allozyme and RAPD variation in spider populations from patchy forest habitats. Genetica 119:27–34.

Hamilton, C.A., B.E. Hendrixson, M.S. Brewer & J.E. Bond. 2014. An evaluation of sampling effects on multiple DNA barcoding methods leads to an integrative approach for delimiting species: a case study of the North American tarantula genus *Aphonopelma* (Araneae, Mygalomorphae, Theraphosidae). Molecular Phylogenetics and Evolution 71:79–93.

Hedin, M.C. 1997a. Molecular phylogenetics at the population/species interface in cave spiders of the southern Appalachians (Araneae: Nesticidae: *Nesticus*). Molecular Biology and Evolution 14:309–324.

Hedin, M.C. 1997b. Speciational history in a diverse clade of habitat-specialized spiders (Araneae: Nesticidae: *Nesticus*): Inferences from geographic-based sampling. Evolution 51:1929–1945.

Hedin, M. & J.E. Bond. 2006. Molecular phylogenetics of the spider infraorder Mygalomorphae using nuclear rRNA genes (18S and 28S): Conflict and agreement with the current system of classification. Molecular Phylogenetics and Evolution 41:454–471.

Hedin, M., S. Derkarabetian, M. McCormack, C. Richart & J.W. Shultz. 2010. The phylogenetic utility of the nuclear protein-coding gene EF-1α for resolving recent divergences in Opiliones, emphasizing intron evolution. Journal of Arachnology 38:9–20.

Hedin, M., J. Starrett, S. Akhter, A.L. Schönhofer & J.W. Shultz. 2012. Phylogenomic resolution of paleozoic divergences in harvestmen (Arachnida, Opiliones) via analysis of next-generation transcriptome data. PLoS ONE 7:e42888.

Herberstein, M.E. & E. Hebets. 2013. Behaviour: Why are spiders good models for research? Pp. 230–251. *In* Spider Research in the 21st Century: Trends and Perspeetives. (D. Penney, ed.). Siri Scientific Press, Manchester, UK.

Hohenlohe, P.A., S. Bassham, P.D. Etter, N. Stiffler, E.A. Johnson & W.A. Cresko. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genetics 6:e1000862.

Hormiga, G., M.A. Arnedo & R.G. Gillespie. 2003. Speciation on a conveyor belt: sequential colonization of the Hawaiian Islands by *Orsonwelles* spiders (Araneae: Linyphiidae). Systematic Biology 52:70–88.

Huber, B.A. 2005. Sexual selection research on spiders: progress and biases. Biological Reviews 80:363–385.

Hudson, P. & M. Adams. 1996. Allozyme characterisation of the salt lake spiders (*Lycosa*: Lycosidae: Araneae) of southern Australia: Systematic and population genetic implications. Australian Journal of Zoology 44:535–567.

Iorizzo, M., D. Senalik, M. Szklarczyk, D. Grzebelus, D. Spooner & P. Simon. 2012. De novo assembly of the carrot mitochondrial genome using next generation sequencing of whole genomic DNA provides first evidence of DNA transfer into an angiosperm plastid genome. BMC Plant Biology 12:61.

Ji, Y.J., H. Smith, D.X. Zhang & G.M. Hewitt. 2004. Ten polymorphic microsatellite DNA loci for paternity and population genetics analysis in the fen raft spider (*Dolomedes plantarius*). Molecular Ecology Notes 4:274–276.

Johannesen, J., T. Baumann, A. Seitz & M. Veith. 1998. The significance of relatedness and gene flow on population genetic structure in the subsocial spider *Eresus cinnaberinus* (Araneae: Eresidae). Biological Journal of the Linnean Soeiety 63:81–98.

Johannesen, J. & Y. Lubin. 1999. Group founding and breeding structure in the subsocial spider *Stegodyphus lineatus* (Eresidae). Heredity 82:677–686.

Johannesen, J. & Y. Lubin. 2001. Evidence for kin-structured group founding and limited juvenile dispersal in the sub-social spider *Stegodyphus lineatus* (Araneae, Eresidae). Journal of Arachnology 29:413–422.

Johannesen, J. & M. Veith. 2001. Population history of *Eresus cinnaberinus* (Araneae: Eresidae) colour variants at a putative species transition. Heredity 87:114–124.

Johannsen, J., A. Hennig, B. Dommermuth & J.M. Schneider. 2002. Mitochondrial DNA distributions indicate colony propagation by single matri-lineages in the social spider *Stegodyphus dumicola* (Eresidae). Biological Journal of the Linnean Society 76:591–600.

Jones, J.S. 1991. Songs in the key of life. Nature 354:323–323.

Jung, J., J.W. Lee, J.P. Kim & W. Kim. 2006. Genetic variations of the golden orb-web spider *Nephila clavata* (Araneae: Tetragnathidae) in Korea, using AFLP markers. Korean Journal of Genetics 28:325–332.

Kanayama, M., Y. Akiyama-Oda & H. Oda. 2010. Early embryonic development in the spider *Achaearanea tepidariorum*: microinjection verifies that cellularization is complete before the blastoderm stage. Arthropod Structure & Development 39:436–445.

King, G.F. & M.C. Hardy. 2013. Spider-venom peptides: structure, pharmacology, and potential for control of insect pests. Annual Review of Entomology 58:475–496.

Klompen, H., M. Lekveishvili & W.C. Black, IV. 2007. Phylogeny of parasitiform mites (Acari) based on rRNA. Molecular Phylogenetics and Evolution 43:936–951.

Koren, S., M.C. Schatz, B.P. Walenz, J. Martin, J.T. Howard & G. Ganapathy, et al. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature Biotechnology 30:693–700.

Krantz, G.W., & D.E. Walter (eds.). 2009. A Manual of Acarology, third ed. Texas Tech University Press, Lubbock, Texas.

Krehenwinkel, H. & D. Tautz. 2013. Northern range expansion of European populations of the wasp spider *Argiope bruennichi* is associated with global warming: eorrelated genetic admixture and population-specific temperature adaptations. Molecular Ecology 22:2232–2248.

Kuntner, M., J.A. Coddington & J.M. Schneider. 2009. Intersexual arms race? Genital coevolution in nephilid spiders (Araneae, Nephilidae). Evolution 63:1451–1463.

Lambeets, K., P. Breyne & D. Bonte. 2010. Spatial genetic variation of a riparian wolf spider *Pardosa agricola* (Thorell 1856) on lowland river banks: the importance of funetional connectivity in linear spatial systems. Biological Conservation 143:660–668.

Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody & J. Baldwin, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lartillot, N. & H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular Biology and Evolution 21:1095–1109.

Lartillot, N., T. Lepage & S. Blanquart. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

Lemmon, A.R., S.A. Emme & E.M. Lemmon. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Systematic Biology 61:727–744.

Liu, L., Y. Li, S. Li, N. Hu, Y. He & R. Pong, et al. 2012. Comparison of next-generation sequencing systems. Journal of Biomedicine and Biotechnology 2012:1–11.

Lohmueller, K.E. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genetics 710:e1002326.

López-Giráldez, F. & J.P. Townsend. 2011. PhyDesign: an online application for profiling phylogenetic informativeness. BMC Evolutionary Biology 11:152.

Masta, S.E. & J.L. Boore. 2004. The complete mitochondrial genome sequence of the spider *Habronattus oregonensis* reveals rearranged and extremely truncated tRNAs. Molecular Biology and Evolution 21:893–902.

Masta, S.E. & J.L. Boore. 2008. Parallel evolution of truncated transfer RNA genes in arachnid mitochondrial genomes. Molecular Biology and Evolution 25:949–959.

Masta, S.E., A.E. Klann & L. Podsiadlowski. 2008. A comparison of the mitochondrial genomes from two families of Solifugae (Arthropoda: Chelicerata): Eremobatidae and Ammotrechidae. Gene 417:35–42.

Masta, S.E., S.J. Longhorn & J.L. Boore. 2009. Arachnid relationships based on mitochondrial genomes: Asymmetric nucleotide and amino acid bias affects phylogenetic analyses. Molecular Phylogenetics and Evolution 50:117–128.

Masta, S.E. 2010. Mitochondrial rRNA secondary structures and genome arrangements distinguish chelicerates: comparisons with a harvestman (Arachnida: Opiliones: *Phalangium opilio*). Gene 449:1–9.

Masella, A.P., A.K. Bartram, J.M. Truszkowski, D.G. Brown & J.D. Neufeld. 2012. PANDAseq: PAired-eND Assembler for Illumina sequences. BMC Bioinformatics 13:31.

Mattila, T.M., J.S. Bechsgaard, T.T. Hansen, M.H. Schierup & T. Bilde. 2012. Orthologous genes identified by transcriptome sequencing in the spider genus *Stegodyphus*. BMC Genomics 13:70.

McCormack, J.E., S.M. Hird, A.J. Zellmer, B.C. Carstens & R.T. Brumfield. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution 62:397–406.

Meyer, M. & M. Kircher. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols, 2010, Cold Spring Harbor, New York, USA.

Meyer, M., U. Stenzel & M. Hofreiter. 2008. Parallel tagged sequencing on the 454 platform. Nature Protocols 3:267–278.

Miller, M.R., J.P. Dunham, A. Amores, W.A. Cresko & E.A. Johnson. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Research 17:240–248.

Mittmann, B. & C. Wolff. 2012. Embryonic development and staging of the cobweb spider *Parasteatoda tepidariorum* C. L. Koch 1841 (syn.: *Achaearanea tepidariorum*; Araneomorphae; Theridiidae). Development Genes and Evolution 222:189–216.

Mortazavi, A., B.A. Williams, K. McCue, L. Schaeffer & B. Wold. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature Methods 5:621–628.

Moya-Laraño, J., M.W. Foellmer, S. Pekár, M.A. Arnedo, T. Bilde & Y. Lubin. 2013. Evolutionary ecology: Linking traits, selective pressures, and ecological functions. Pp. 112–153. *In* Spider Research in the 21st Century: Trends and Perspectives. (D. Penney, ed.). Siri Scientific Press, Manchester, UK.

Nielsen, R. 2010. Genomics in search of rare human variants. Nature 467:1050–1051.

Osakabi, M.H. & Y. Sakagami. 1994. RFLP analysis of ribosomal DNA in sibling species of spider mite, genus *Panonychus* (Acari: Tetranychidae). Insect Molecular Biology 3:63–66.

Ovchinnikov, S. & S.E. Masta. 2012. Pseudoscorpion mitochondria show rearranged genes and genome-wide reductions of RNA gene sizes and inferred structures, yet typical nucleotide composition bias. BMC Evolutionary Biology 12:31.

Oxford, G.S. 2005. Genetic drift within a protected polymorphism: Enigmatic variation in color-morph frequencies in the candy-stripe spider, *Enoplognatha ovata*. Evolution 59:2170–2184.

Oxford, G.S. & B. Gunnarsson. 2006. Spatial variation in colour morph, spotting and allozyme frequencies in the candy-stripe spider, *Enoplognatha ovata* (Theridiidae) on two Swedish archipelagos. Genetica 128:51–62.

Parmakelis, A., K. Balanika, S. Terzopoulou, F. Rigal, R.R. Beasley & K.L. Jones, et al. 2013. Development of 28 polymorphic microsatellite markers for the endemic Azorean spider *Sancus acoreensis* (Araneae, Tetragnathidae). Conservation Genetics Resources 5:1133–1134.

Pedersen, A.A. & V. Loeschcke. 2001. Conservation genetics of peripheral populations of the mygalomorph spider *Atypus affinis* (Atypidae) in northern Europe. Molecular Ecology 10:1133–1142.

Pepato, A.R., C.E. da Rocha & J.A. Dunlop. 2010. Phylogenetic position of the acariform mites: sensitivity to homology assessment under total evidence. BMC Evolutionary Biology 10:235.

Peterson, B.K., J.N. Weber, E.H. Kay, H.S. Fisher & H.E. Hoekstra. 2012. Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. PLoS ONE 7:e37135.

Piel, W.H. & K.J. Nutt. 2000. One species or several? Discordant patterns of geographic variation between allozymes and mtDNA sequences among spiders in the genus *Metepeira* (Araneae: Araneidae). Molecular Phylogenetics and Evolution 15:414–418.

Pond, S.L.K., S.D.W. Frost & S.V. Muse. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679.

Pons, J. & R.G. Gillespie. 2003. Common origin of the satellite DNAs of the Hawaiian spiders of the genus *Tetragnatha*: evolutionary constraints on the length and nucleotide composition of the repeats. Gene 313:169–177.

Pons, J. & R.G. Gillespie. 2004. Evolution of satellite DNAs in a radiation of endemic Hawaiian spiders: Does concerted evolution of highly repetitive sequences reflect evolutionary history? Journal of Molecular Evolution 59:632–641.

Porter, A.H. & E.M. Jakob. 1990. Allozyme variation in the introduced spider *Holocnemus Pluchei* (Araneae, Pholcidae) in California. Journal of Arachnology 18:313–319.

Porter, M.L., T.W. Cronin, D.A. McClellan & K.A. Crandall. 2006. Molecular characterization of crustacean visual pigments and the evolution of pancrustacean opsins. Molecular Biology and Evolution 24:253–268.

Prendini, L. & L.A. Esposito. 2010. A reanalysis of *Parabuthus* (Scorpiones: Buthidae) phylogeny with descriptions of two new *Parabuthus* species endemic to the Central Namib gravel plains, Namibia. Zoological Journal of the Linnean Society 159:673–710.

Price, M.N., P.S. Dehal & A.P. Arkin. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE 5:e9490.

Quail, M.A., M. Smith, P. Coupland, T.D. Otto, S.R. Harris & T.R. Connor, et al. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341.

Ramirez, M.G. & K.E. Haakonsen. 1999. Gene flow among habitat patches on a fragmented landscape in the spider *Argiope trifasciata* (Araneae: Araneidae). Heredity 83:580–585.

Ramirez, M.G. & B. Chi. 2004. Cryptic speciation, genetic diversity and gene flow in the California turret spider *Atypoides riversi* (Araneae: Antrodiaetidae). Biological Journal of the Linnean Society 82:27–37.

Rasmussen, M, et al. 2011. An aboriginal Australian genome reveals separate human dispersals into Asia. Seience 334:94–98.

Reed, D.H., A.C. Nicholas & G.E. Stratton. 2007. Inbreeding levels and prey abundance interact to determine fecundity in natural populations of two species of wolf spider. Conservation Genetics 8:1061–1071.

Reed, D.H., V.H. Teoh, G.E. Stratton & R.A. Hataway. 2011. Levels of gene flow among populations of a wolf spider in a recently fragmented habitat: current versus historical rates. Conservation Genetics 12:331–335.

Robinson, M.D., D.J. McCarthy & G.K. Smyth. 2009. edgeR: a Bioconductor package for differential cxpression analysis of digital gene expression data. Bioinformatics 26:139–140.

Rocha, L.A., M.A. Bernal, M.R. Gaither & M.E. Alfaro. 2013. Massively parallel DNA sequeneing: the new frontier in biogeography. Frontiers of Biogeography 5.1.

Rothberg, J.M., W. Hinz, T.M. Reariek, J. Schultz, W. Mileski & M. Davey, et al. 2012. An integrated semieonductor device enabling non-optical genome sequencing. Nature 475:348–352.

Rota-Stabelli, O., E. Kayal, D. Gleeson, J. Daub, J.L. Boore & M.J. Telford, et al. 2010. Ecdysozoan mitogenomics: evidenee for a common origin of the legged invertebrates, the Panarthropoda. Genome Biology and Evolution 2:425–440.

Rutten, K.B., I. Schulz, K. Olek & G. Uhl. 2001. Polymorphic microsatellite markers in the spider *Pholcus phalangioides* isolated from a library enriched for CA repeats. Molecular Eeology Notes 1:255–257.

Salomone, N., V. Vignoli, F. Frati & F. Bernini. 2007. Species boundaries and phylogeography of the "*Euscorpius carpathicus* complex" (Scorpiones: Euscorpiidae) in Italy. Molecular Phylogenetics and Evolution 43:502–514.

Satler, J.D., B.C. Carstens & M. Hedin. 2013. Multilocus species delimitation in a complex of morphologically conserved trapdoor spiders (Mygalomorphae, Antrodiaetidae, *Aliatypus*). Systematic Biology 62:805–823.

Schafer, M.A., A. Hille & G.B. Uhl. 2001. Geographical patterns of genetic subdivision in the cellar spider *Pholcus phalangioides* (Araneae). Heredity 86:94–102.

Sehafer, M.A. & G. Uhl. 2002. Determinants of paternity suecess in the spider *Pholcus phalangioides* (Pholcidae: Araneae): the role of male and female mating behaviour. Behavioral Ecology and Sociobiology 51:368–377.

Shendure, J. & H. Ji. 2008. Next-generation DNA sequencing. Nature Bioteehnology 26:1135–1145.

Smith, D.R. & M.S. Engel. 1994. Population structure in an Indian cooperative spider, *Stegodyphus sarasinorum* Karsch (Eresidae). Journal of Arachnology 22:108–113.

Smith, D.R. & R.H. Hagen. 1996. Population structure and interdemic selection in the cooperative spider *Anelosimus eximius*. Journal of Evolutionary Biology 9:589–608.

Smith, D.R. 2013. RNA-Seq data: a goldmine for organelle research. Briefings in Functional Genomics 12:454–456.

Stamatakis, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Stapley, J., J. Reger, P.G.D. Feulner, C. Smadja, J. Galindo & R. Ekblom, et al. 2010. Adaptation genomics: the next generation. Trends in Ecology and Evolution 25:705–712.

Starrett, J. & M. Hedin. 2007. Multilocus genealogies reveal multiple cryptie species and biogeographical complexity in the California turret spider *Antrodiaetus riversi* (Mygalomorphae, Antrodiaetidae). Molecular Ecology 16:583–604.

Steiner, W.W.M., M.H. Greenstone & G.E. Stratton. 1992. Variation in *Schizocosa* (Araneae: Lycosidae), *Metaphidippus* and *Phidippus* (Araneae: Salticidae). Journal of Arachnology 20:35–39.

Su, Y.C., Y.H. Chang, D. Smith, M.S. Zhu, M. Kuntner & I.M. Tso. 2011. Biogeography and speciation patterns of the golden orb spider genus *Nephila* (Araneae: Nephilidae) in Asia. Zoological Science 28:47–55.

The Honeybee Genome Sequencing Consortium. 2006. Insights into soeial insects from the genome of the honeybee *Apis mellifera*. Nature 443:931–949.

The International Aphid Genomics Consortium. 2010. Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biol 8:e1000313.

Tso, I.M., P.L. Tai, T.H. Ku, C.H. Kuo & E.C. Yang. 2002. Colour-associated foraging success and population genetic structure in a sit-and-wait predator *Nephila maculata* (Araneae: Tetragnathidae). Animal Behaviour 63:175–182.

Valdez-Velázquez, L.L., V. Quintero-Hernández, M.T. Romero-Gutiérrez, F.I.V. Coronas & L.D. Possani. 2013. Mass fingerprinting of the venom and transcriptome of venom gland of scorpion Centruroides tecomanus. PLoS ONE 8:e66486.

Van Tassell, C.P., T.P. Smith, L.K. Matukumalli, J.F. Taylor, R.D. Schnabel & C.T. Lawley, et al. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods 5:247–252.

Vandergast, A.G., R.G. Gillespie & G.K. Roderick. 2004. Influence of volcanic activity on the population genetic structure of Hawaiian *Tetragnatha* spiders: fragmentation, rapid population growth and the potential for accelerated evolution. Molecular Ecology 13:1729–1743.

Viljakainen, L., J.D. Evans, M. Hasselmann, O. Rueppell, S. Tingek & P. Pamilo. 2009. Rapid evolution of immune proteins in social insects. Molecular Biology and Evolution 26:1791–1801.

Waterhouse, R.M., F. Tegenfeldt, J. Li, E.M. Zdobnov & E.V. Kriventseva. 2012. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Research 41:D358–D365.

Wolff, C. & M. Hilbrant. 2011. The embryonic development of the Central American wandering spider *Cupiennius salei*. Frontiers in Zoology 8:15.

Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24:1586–1591.

Yip, E.C., D.M. Rowell & L.S. Rayor. 2012. Behavioural and molecular evidence for selective immigration and group regulation in the social huntsman spider, *Delena cancerides*. Biological Journal of the Linnean Society 106:749–762.

Zhulidov, P.A., E.A. Bogdanova, A.S. Shcheglov, I.A. Shagina, L.L. Wagner & G.L. Khazpekov, et al. 2005. A method for the preparation of normalized cDNA libraries enriched with full length sequences. Russian Journal of Bioorganic Chemistry 31:170–177.

Zobel-Thropp, P.A., M.R. Bodner & G.J. Binford. 2010. Comparative analyses of venoms from American and African *Sicarius* spiders that differ in sphingomyelinase D activity. Toxieon 55:1274–1282.

Manuscript received 3 October 2013, revised 7 January 2014.